What Corpus Linguistics can offer Contact Linguistics: the C-ORAL-BRASIL corpus experience

O que a Linguística de Corpus *pode oferecer à Linguística de Contato: a experiência do* corpus C-ORAL-BRASIL

Heliana Mello
Universidade Federal de Minas Gerais, Brasil
hmello@ufmg.br

Abstract: Contact Linguistics, throughout its history, has been mostly a data-oriented subdiscipline. From the gathering of word lists in colonial settings by pioneer scholars to the current compilation of narratives, interviews and databanks, Contact Linguistics, differently from other Linguistics subdisciplines, has strived to base its findings on the analysis of actual language produced by speakers of a given language, and not on solely introspective methodologies. On the other hand, Corpus Linguistics has brought innovative methodological approaches to mostly every subfield in linguistic research. Corpora compilation parameters have taken representativeness and balance seriously and this, in turn, has aided the finding of generalizations about linguistic systems. In this paper, the C-ORAL-BRASIL corpus, a spontaneous speech corpus of informal Brazilian Portuguese is presented, and some of its characteristics that may be explored very profitably by a Contact Linguistics perspective are highlighted.

Keywords: Contact Linguistics, corpus linguistics, spontaneous speech, C-ORAL-BRASIL

Resumo: Ao longo de sua história, a Linguística de Contato desenvolveu-se, majoritariamente, como uma disciplina baseada em dados. Isso se nota através da coleta de listas de palavras, em séculos passados, até as atuais coletas de narrativas, entrevistas e formação de bancos de dados. A Linguística de Contato, diferentemente de outros campos disciplinares da Linguística, manteve-se fiel ao embasamento de descobertas a partir de análises linguísticas baseadas em dados produzidos por falantes de uma dada língua, e não apenas por metodologias introspectivas. Por outro lado, a Linguística de Corpus tem trazido metodologias inovadoras para um amplo espectro de campos investigativos da Linguística. Parâmetros de compilação de corpora levam seriamente em contra noções como representatividade e balanceamento de dados e isto, por sua vez, tem auxiliado a descoberta de importantes generalizações sobre os sistemas linguísticos. Neste artigo, o C-ORAL-BRASIL, um corpus de fala espontânea informal do português brasileiro é apresentado e algumas de suas características, que podem ser exploradas sob o ponto de vista da Linguística de Contato, são destacadas.

Palavras-chave: Linguística de contato, linguística de corpus, fala espontânea, C-ORAL-BRASIL.

1 Introduction

Language contact, variation and change have been a topic of interest for many a researcher for a long time. I was introduced to this area in the early 1990s by having the privilege of being advised in my doctoral thesis by Prof. John Holm, at the Graduate Center at the City University of New York. The program I was enrolled at had a very strong lean towards formal linguistic theories and introspective methods of language analysis at the time. Prof. Holm captured my attention and respect for several reasons, but not least his strong standing towards actual data analysis to support linguistic findings and a keen interest in language contact, including what at the time he formulated as semicreolization — this process, in his view, would characterize the genesis of somewhat mixed languages, which encompassed my mother tongue, Brazilian Portuguese, as a highly likely candidate.

In the early 1990's I carried fieldwork in the conventional way it was done at the time, much like what Prof. Holm had described as his own experience years earlier in Nicaragua: I visited the community I was interested in studying, proceeded to record as many situations and as many people as possible, focused on narratives and interviews, particularly those that I could obtain from elderly speakers. The next step in this process was the painstaking transcription of low quality audio files and the manual search for interesting patterns of all sorts: lexical, semantic, morphological, syntactic and even phonetic.

Two decades have passed since then — and the technology that is readily available to language studies today seems to be a world apart from those early graduate school days. This is what I am going to focus on in this paper — specifically, what the field of corpus linguistics can offer language contact studies, through the description of a spontaneous speech corpus of Brazilian Portuguese: C-ORAL-BRASIL.

2 Corpus Linguistics and Contact Linguistics

Contact linguistics is an interdisciplinary area of investigation that draws from different linguistic subfields, such as second language acquisition, sociolinguistics and theoretical linguistics, among others, besides taking into account historical and cultural theories. The field is sometimes classified as a branch of sociolinguistics but has been gaining ground as a specific domain of inquiry in the past two decades (cf. Hickey 2012, Matras 2009, Holm 2003, Winford 2003, Thomason 2001).

Contact linguistics focuses on issues such as linguistic borrowing, mixing, emergence of new language varieties such as pidgins and creoles, language shift, linguistic hegemony, among others. As globalization and population movements have grown, so has language contact. Therefore, this has been an area of increasing interest, not least to inform language policies in multilingual areas of the world (Hornberger 2002).

The study of contact phenomena in all their linguistic analytical levels depends on empirical evidence provided by actual data. This is the very premise for sound contact linguistics that can be enhanced through corpus linguistic methodologies.

Corpus linguistics is devoted to the compilation, annotation and study of written, spoken and multimedia electronic corpora. The history of corpus linguistics goes back decades, if not centuries, when scholars initially manually gathered and analyzed data, such is said to be the case of J. Kading, a German linguist, who analyzed 11 million words in 1897 (Meyer 2008). However, in more recent linguistic history, electronic corpus linguistics has been said to have sprouted to the mainstream with Francis and Kucera who launched the first electronic corpus — the Brown Corpus in its first version in 1964 (http://khnt.aksis.uib.no/icame/manuals/brown/).

Corpus linguistics has been growing steadily for some thirty years and has been used in a large number of linguistic applications and studies, majorly in lexicography, grammar, translation, stylistics, and second language studies.

Although most well-known corpora are still written corpora, spoken corpora, as well as multimedia corpora, have been increasingly compiled.

Contact linguistics, with its interest in phenomena which emerge from the inter-speaker contact of languages in multilingual environments, by definition, is more akin to studying spoken than written data. The written codification of contact varieties is of course possible (cf. standardization of written codes for creole languages) and is becoming a subject of interest on its own right. However, it is evident that contact among languages first takes place in the spoken realm, in which its effects are felt most immediately, only to transit into written codification much later. Thus, it is only natural that spoken corpora compilation might be an immensurable source of support for language contact studies.

Shifting back to my graduate school days, I carried research on the genesis and development of Brazilian vernacular Portuguese (Mello 1997), focusing particularly on contact phenomena and exploring the creolization hypothesis under Prof. John Holm's supervision. Although some of the data I studied then came from 18th and 19th century travelers' notes and novels, the absolute majority of my analysis was carried taking spoken data into consideration. Therefore, I then focused on the lexicon, grammar (syntax, semantics, phonology) and some segmental phonetic aspects of Brazilian Portuguese. An issue I did not explore in my thesis was prosodic effects of language contact and possible pragmatic entailments ensued by contact. At the time I had hopes that in the future there would be the possibility of studying prosody as a major element in language contact with all its consequences, phonological, but equally relevant, pragmatic.

The third generation of corpora allows us to do just what I had hoped for in the 1990s. Spoken corpora that bring not only transcription and sound files, but actually make available speech-to-text alignment, offer the possibility of comparative studies focused on prosodic features related to pragmatic aspects, majorly the organization of information structure that just might shed a new light in the direction of contact linguistics. All other analytical levels are entirely possible through these corpora as well; therefore more traditional lines of study, such as syntax and lexicon, can be explored very profitably; which, as a matter of fact, can be much more accurately dealt with methodologically through computational techniques than manual ones.

There are not many contact varieties corpora freely available for study at this stage, which can be consulted online or under request[1]. Among the corpora (and studies derived from them) available or in the process of compilation are: Corpus of Northern Haitian Creole, a spoken corpus made available by Indiana University (http://www.indiana.edu/˜creole/), Corpus of Written British Creole made available by Lancaster University (www.ling.lancs.ac.uk/staff/mark/cwbc/cwbcman.htm) and Corpus of Santomense, still in the making, under the responsibility of Tjerk Hagemeijer and associates (Hagemeijer *et al.* 2012) at CLUL in Lisbon. Relevant to the current discussion are papers dealing with statistical approaches to contact varieties such as Robinson (2008) and Siegel *et al.* (in press). Still of interest are computational approaches to contact varieties such as Mason and Allen (2003).

In section 3, the C-ORAL-BRASIL, an informal spontaneous speech Brazilian Portuguese corpus is thoroughly discussed, with the purpose of showing how such a tool could be very favorably used in language contact-based studies.

3 The C-ORAL-BRASIL

The C-ORAL-BRASIL (Raso & Mello 2012)[2] is a Brazilian Portuguese spontaneous speech corpus, especially representative of the Mineiro diatopy, majorly from the metropolitan region of the state capital Belo Horizonte. The corpus comprises about 20 hs of recording, around 208,000 words, and 139 texts. The corpus features private/familiar and public contexts, through monologues, dialogues and conversations. Some of the diaphasic variations recorded were, for example, a soccer match, a construction worker and an engineer at a building site, drag-queens putting make up on before a show, waiters waiting at a party, among several others. The texts were recorded with sophisticated wireless equipment, in order to guarantee highly accurate acoustic quality, between 2006 and 2011. C-ORAL-BRASIL is structured to be

[1] I thank an anonymous referee for pointing out the Cocoon interface (cocoon.tge-adonis.fr) that offers free access to a large spoken database that includes recordings of creole speakers. This, however, is not a corpus resource in the sense a corpus has been defined in this paper. Equally relevant was the referee's mention of Dominique Fattier's seminal work on Haitian Creole varieties which resulted in the *L'atlas d'Haïti*. This again, although based on very sound data collecting, is not based on an electronic corpus.

[2] The C-ORAL-BRASIL Project was funded by Fapemig, CNPq and UFMG.

comparable with the C-ORAL-ROM project corpora (Cresti & Moneglia 2005)[3] for French, Italian, Spanish and European Portuguese. Here, the central information about the corpus and the motivation for its architecture and sampling methodology are listed, in an attempt to show the advantages that they present for the study of spontaneous speech in its different levels, which include prosodic and pragmatic dimensions.

The corpus DVD contains:

1. the multimedia corpus, made up of the following archives for each text: audio (wav), transcription (rtf) and aligned file (xml) through WinPitch software (Martin 2005), and txt file;
2. the metadata: title, file name, participant abbreviation and their main sociolinguistic characteristics (gender, age, school level, occupation and role played in the interaction), recording date, place, context and topic, corpus branch, duration in time and number of words, acoustic quality, transcribers and revisers' names, and any commentary considered useful;
3. the corpus tagged lexically and morphosyntactically (Bick 2012) in full version (xml and txt) and in a simplified version (xml and txt);
4. frequency lists, spreadsheets with relevant linguistic measurements and statistics about the informants;
5. a book, in pdf format in which audio examples are linked to the text, containing the corpus description, a presentation of the theory behind it, the explanation for transcription and segmentation criteria and their validation, a discussion of the main speech measurements, and finally a description and discussion of the parser used for the lexical and morphosyntactic tagging.

The corpus transcription format follows CHAT (MacWhinney 2000), implemented for prosodic annotation (Moneglia & Cresti 1997); the corpus is segmented into utterances and tone units (Raso 2012b; Mello *et al.* 2012). The utterance is defined as the minimal unit with pragmatic autonomy. Its identification is marked by a prosodic break perceivable as terminal. This is illustrated through (1). The linguistic sequence in (1a) can, in principle, be segmented in different ways. A simple reading induces the interpretation of *mas os filho também nũ são fácil também* as an autonomous entity since it is syntactically autonomous (one sentence); the rest can be also interpreted as one or more entities. Nevertheless, by listening to the sequence, it is clear that there is just one autonomous entity, that is, one utterance, segmentable as follows (1b):

[3] For a comparison between C-ORAL-BRASIL and C-ORAL-ROM, see Raso (2012a); Mittmann and Raso (2012).

(1) (*bfammn*03)[4]

a. *ALO: *mas os filho também nũ são fácil também juntou os filho todo foram lá e trouxeram o corpo na força* [but the sons too they are not easy either they all meet (they) go there and bring the body by force]

b. *ALO: *mas os filho também nũ são fácil também / juntou os filho todo / foram lá e trouxeram o corpo na força //*

The double slash marks a terminal break, that is, the utterance frontier; the single slash marks a tone unit frontier. In fact, the first part of the sequence, which could seem autonomous in reading, is not perceived as such through listening to the actual recording.

Examples (2) and (3) show that the same syntactic structure (in both cases a main clause followed by a relative clause) can be the locutionary content of one or more than one utterances, depending on their prosodic realization:

(2) (*bfamdl*02)
*BAL: *tá saindo de uma garrafinha que tem um bico muito pequeno //*
[It's coming out from a little bottle with a very small neck]

(3) (*bfamdl*02)
eê tá com um jarro d´água // que tem uma espessura assim //
[you have a water jar that has a width like this]

Apparently in (3) we have only one utterance. But listening to the sequence we realize that it performs two autonomous utterances.

Example (4) shows that a linguistic sequence that could be interpreted by a reader as a negative assertion, if listened to, is clearly a sequence in which an affirmative utterance is preceded by another utterance that expresses refusal:

(4) (*bpubdl*01)
*PAU: *não // tá dando a altura daquele que a Isa marcou lá / né //*
[no // it has the height of that one that Isa marked there / isn't it //], also interpretable by the reader as: [it doesn't have the height of that one that Isa marked there / does it //]

[4] All the examples cited in this paper can be listened to in the C-ORAL-BRASIL DVD.

Example (5) and Figure 1 show that the terminal break does not necessarily match with a pause. As the formants in the figure show, there is no pause between the first and the second utterances, while utterance two and three are divided by a pause:

(5) (_bfamdl_02)
*BAL: t_á saindo de uma garrafinha que tem um bico muito pequeno // então daquela coisa pequeninim nũ vai encher rápido // agora imagina cê pega um balde e joga dentro //_
[It's coming out from a little bottle with a very small neck // so that little thing can't fill it quickly // now you imagine you fill it with a full bucket //]

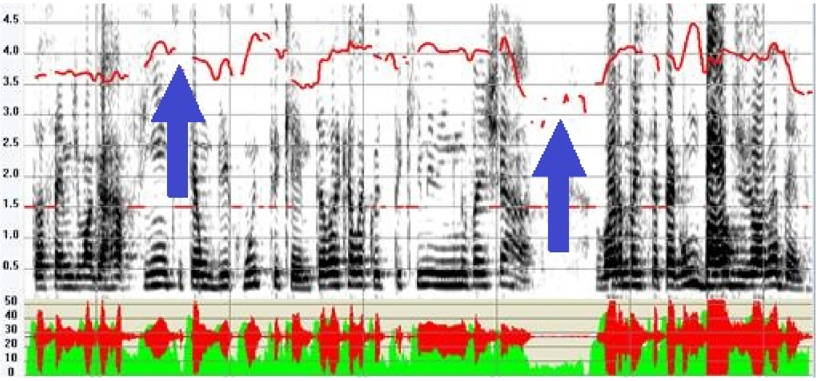

Fig. 1: Example (5) represented in WinPitch software.

The opposite is also true: a pause, even a long one, does not imply an utterance frontier, as in example (6), in which we have a pause of 1,281 ms inside the utterance:

(6) (_bpubdl_11)
*MAR: _o ensino tá [/1] tá assim / difícil / mas tá mais fácil / né hhh //_
[teaching is how can I say difficult but easier]

Therefore, only by listening to a verbal sequence is it feasible to understand where a pragmatically interpretable reference unit ends. Hence, it is not possible to analyze speech without audio, nor is it possible to transcribe speech without marking the reference units that are the basis for its segmentation.

That cannot be perceived by reading, nor be automatically measured through a pause (Moneglia 2005).

These are the main reasons why the C-ORAL-BRASIL is sound-text aligned for each utterance. Alignment is a crucial aspect in the study of speech. Without sound alignment, the text cannot be appropriately studied, since the audio source turns out to be unusable and unrecoverable for research. Without resort to the audio and to text/audio alignment the real object of study, would be the transcription, which represents a special variety of writing, without the basic characteristics of speech, above all prosody. In our view, it is not possible to study speech without its acoustic information, that alone allows for the recognition of the main categories that structure speech and deem it interpretable, illocution being the basic one (Cresti 2000b). In fact, example (7) shows that a pure syntactic and semantic analysis that does not result from an illocutive interpretation cannot account for the understanding of speech:

(7) (*bfamdl*04)
*KAT: *o quê*
[what //]
*SIL: *copos // copos de Urano / que tem aí //* [glasses // glasses from Urano / that are here //]
*KAT: *copos de quê //* [glasses made of what //]
*SIL: *Urano //* [Urano //]
*KAT: *Urano //* [Urano //]
*SIL: *é // Urano // Urano //* [yeah // Urano // Urano]

In (7) the different occurrences of *Urano // Urano* stand for respectively: a confirmation, expression of incredulity, and the last two are conclusions, uttered with different attitudes. It is only through the different illocutions that we can recover the different meanings of *Urano* in the different utterances. Its communicative function cannot be recovered simply through its semantic and syntactic forms.

3.1 The architecture

By spontaneous speech is meant speech that is planned at the same moment it is performed, i.e. speech that does not perform a previously, totally or partially, planned text, like acted speech or even a previously planned discourse (Nencioni 1983; Cresti 2000a; Biber 1988; Blanche-Benveniste *et al.* 1990; Miller & Weinert 1998; Moneglia 2005, 2011). Spoken events that can be considered spontaneous show: i. a multimodal face-to-face interaction; ii. intersubjective reference to the deictic space; iii. mental programming at the same time as vocal performance; iv. contextually undetermined linguistic behavior, i.e. unforeseen behavior.

A long tradition of sociolinguistic studies (Berruto 1987; Biber & Conrad 2001; Biber *et al.* 1998; Gadet 1996a, 1996b, 1997, 2000, 2003; Halliday 1989) focused on the value of sociological and contextual parameters to define speech qualities, and pointed to their variability. There are many types of spontaneous speech, and they vary according to the following parameters: a) the possible structural varieties of the communicative event (monologue, dialogue, conversation); b) the communicative channel; c) the sociological context, that is, the social domain of the event (family, private, public life); d) the programming conditions (partially or totally programmed versus non programmed speech); e) possible register and genre varieties; f) sociolinguistic factors (gender, age, school level, speaker's occupation); g) geographic origin; h) speech event task; i) topic.

Planning a spoken corpus is, therefore, a complex task that must ensure representativity of the principal variations explored by the different types of events in spontaneous speech (Berruto 1987; Biber 1988; De Mauro *et al.* 1993; Gadet 1996a, 1996b, 2003). Most speech resources built so far, usually having technology needs as their objective (telephone information, health interactions, map tasking), were produced in controlled situations. This allows a very high acoustic quality, but represents restrict semantic domains, with highly foreseeable linguistic behavior. On the other hand, C-ORAL-BRASIL, like C-ORAL-ROM, collects data in natural context, which necessarily reduces acoustic quality and causes many more difficulties for recording. C-ORAL-BRASIL underwent a great effort to obtain the best acoustic quality for recordings in very different contexts, using sophisticated wireless equipment.

As previously mentioned, an important goal of this corpus is to achieve comparability with the C-ORAL-ROM corpora. Comparable corpora have been built for written language, mostly parallel corpora or corpora of the same specialized topic. Usually comparable spoken corpora encompass reading tasks which lead to a loss of spontaneity — this is what is usually done in phonetic laboratories. In speech, comparability can easily be reached only in strongly controlled situations. But if we assume that spontaneous speech is necessarily documented maximizing textual variation, the consequence is that the more textual variation we have, the less comparability we obtain. Therefore, comparability among the corpora of the C-ORAL projects results from the application of the same specific compilation parameters, not through the comparability of uttered texts *per se*.

The first half of the C-ORAL-BRASIL corpus, published in 2012, encompasses the informal part of the project. The formal part is still being compiled. The informal corpus features 208,130 words, distributed in 139 texts of, on average, 1,500 words each. A few texts are bigger (up to 5,000 words) or smaller (only if they maintain textual autonomy). The 139 texts were divided in two contexts: private/familiar (159,364 words) and public (48,766 words); for each context the texts were divided similarly among three interactional typologies: monologues, dialogues and conversations (dialogic texts with more than two main participants).

Texts are transcribed using the CHILDES-CLAN format (MacWhinney 2000) implemented for prosodic annotation (Moneglia & Cresti 1997). The prosodic annotation features the segmentation of the speech flow in utterances (double slash) and tone units (single slash)[5]; interrupted utterances (+) and retracting ([/n])[6] are also marked. Transcriptions follow traditional orthography, with significant exceptions due to the goal of capturing speech phenomena that can show processes of gramaticalizations and lexicalizations going on, so that they can be computed and statistically studied[7].

3.2 The pragmatic perspective and diaphasic variation

A truly spontaneous speech corpus must portray in the best possible way situational variation. In fact, the factor which conditions speech structuring the most is not speakers or topic variation. Especially under a pragmatic perspective, it is crucial to document the differences in verbal behavior depending on the different tasks speakers should perform in different situations. If on one hand the sociolinguistic tradition allows us to identify the main domains of formal speech, on the other, the possible situations in informal speech cannot be categorized. Therefore, while in formal speech it is possible to list a certain group of typical contexts, informal variation must be left open. The goal is, therefore, to document the widest range of situations, as no specific context can be considered, in principle, more typical than another.

[5] For the segmentation theoretical frame, see Cresti (2000a); for the segmentation and validation methodology, see Mello *et al.* (2012), Moneglia *et al.* (2010), Raso and Mittmann (2009).

[6] The number means the quantity of retracted words.

[7] For the transcription criteria, see Mello and Raso (2009) and Mello *et al.* (2012).

In order for this to be possible, considered that the cost (both economic and especially concerning time for recordings and transcriptions) of a spoken corpus is much higher than that for written ones, it is important to provide many different texts and reduce their size. The average of 1,500 words is sufficient for the interaction to be autonomous, since this size oral text usually exhibits syntactic and pragmatic properties (Blanche-Benveniste *et al.* 1990; Scarano 2003), and at the same time allows the representation of a wide variety of situations.

Within the informal register, the partition in private/familiar versus public context documents the role that the participant plays, whether he/she acts as an individual, as for example in interactions with relatives or friends, or in a professional or institutional role, as, for example, in interactions between client and sales representative, or student and professor, or citizen and public officer, etc. Around 75% of the corpus represents the private/familiar context, since it normally occupies a larger space in human natural interactions.

Inside each context, there are three different typologies of interactions: i) a monologic typology, in which a speaker builds a spoken text, (almost) without any interaction; ii) a dialogic typology, in which two interlocutors interact; iii) a conversational typology, in which three or more speakers interact. The text characteristics are strongly conditioned by the interaction typology, especially in the opposition between monologic versus interactional[8]. It must be highlighted, however, that, differently from formal, informal register does not exhibit, in principle, perfect monologic texts. Almost always there will be some kind of interaction. The criterion used to assign a text to this typology was the fact that the construction of a spoken text keeps developing even after the interlocutor's interventions which, in the majority of cases, are not considered by the speaker. The monologic typology is built by long turns and within them, by very articulated utterances with complex information structure and many tone units, stretching in a strongly processual way. The reference to the situational context is usually poor, while a great amount of cognitive contextualization is necessary. Depending on the textual typology of the monologic text, the more frequent illocutions change, but the illocutionary variation is poor. On the other hand, interactional typologies show short turns and small informationally patterned utterances; in these cases, the reference to the situational context is strong, making a high amount of verbal contextualization not necessary, while the illocutionary variation inside the same text is very high[9].

[8] See Raso and Mittmann (2012).

[9] See Raso (2012b) and Mittmann and Raso (2012).

After this important distinction between monologic and interactive typologies, the most important factor of variation is dependent upon each typology. In monologic typology, speech structure depends mainly on textual genre: life tale, professional explanation, argumentation, joke, recipe, story, etc. In dialogues and conversations, variation is basically due to the task speakers are performing: a chat between friends is much different from a couple's quarrel, or from an interaction between seller and client, or among the players in a football game, or between a personal trainer and an athlete, or between mother and crying child, or between two interactants performing a task together, etc. It is evident that in each activity the actions to be performed change completely, as well as the turn size, the amount of silence, etc.

These observations should be sufficient to illustrate how crucial the importance of true diaphasic variation is in a spontaneous speech corpus. Speech structuring variation cannot be documented through speakers' or topic variation. Different speakers perform the same action in basically the same way, and the change in topic in chats or interviews does not lead to structural variation, i.e. illocutionary and information structure variation (Cresti 2000b).

3.3 Diastratic variation

Although diaphasic variation had been the main goal while building the corpus architecture, diastratic variation is very well documented. What is important methodologically is that, while diaphasic variation has no chance to be documented aiming only for diastratic variation, our methodology shows that diastratic variation is a natural consequence of the diaphasic one (Cresti & Moneglia 2012).

C-ORAL-BRASIL features 362 speakers. For 68.23% of them gender, age, origin and school level are documented. Although more than 30% of the speakers are not fully documented, this is due to the fact that they entered the recording context in an unforeseen way, without any kind of preparation, as it is the case in naturally occurring interactions. This strongly reinforces the point that the recording context was truly uncontrolled and portrays spontaneous speech. Moreover, they are responsible for only 1.91% of the overall number of corpus words. The distribution of number of words uttered by number of speakers is shown in table 1.

Table 1 shows that 44.5% of the speakers utter up to 247 words, accounting for just 3.92% of the corpus. Table 2 shows the cluster grouping these 247 speakers. Table 2 shows that the great majority of the non-documented speakers (109) utters up to 47 words, and that more than half of them (82)

Tab. 1: Number of words per number of speakers.

Number of words	Number of speakers
1 - 247 words	161
280 - 627 words	81
649 - 908 words	37
933 - 1016 words	16
1134 - 1400 words	26
1455 - 1663 words	17
1777 - 1994 words	7
2140 - 2455 words	10
2611 - 2901 words	2
3550 - 3738 words	2
4211 - 4327 words	2
6309 words	1
TOTAL	362

Tab. 2: Grouping of speakers uttering up to 247 words.

Word number clusters	speakers
1 - 22 words	82
25 - 47 words	27
54 - 72 words	10
77 - 95 words	13
99 - 115 words	6
136 - 164 words	9
172 - 185 words	5
204 - 247 words	8
TOTAL	161

utter up to 22 words. Table 1 shows also that the corpus features a small group of speakers (5) that utter more than 3,550 words each, representing 10.63% of the overall number of words. These speakers appear in more than one recording in different situations and may be studied to see how the same speaker's speech varies in different contexts.

Gender balancing is very even in terms of number of words: 50.36% of the speakers are female and 49.64% are male. In terms of speakers, 203 are female and 159 male (one informant utters just one word and his/her gender was not identified). Age balancing is also very good (measurements in words): 27.13% of the speakers belong to group A (from 18 to 25 years old); 30.28% to group B (from 26 to 40 years old); 31,01% to group C (41 to 60 years old); 8,05% to group D (more than 60 years old); 1.61% are underage and 1.91% are not documented for age (group M). The corpus is very well balanced as far as speakers older than 18 years are concerned, considering that group D in Brazilian society is smaller than the others. As far as the number of speakers is concerned, 75 are in group A, 1 is registered in group A in one interaction and in group B later, 88 belong to group B, 64 to group C, 15 to group D and 11 to group M.

Schooling is very well represented for mid and high schooling levels, the most relevant in the representation of the language synchronic standard use, but low schooling level is also sufficiently represented. Taking word numbers, 15.79% represent level 1 (no more than 7 years of school), 40.76% represent level 2 (up to college degree if the degree is never used for their occupation), 40.66% represent level 3 (use college degree for their occupation or have degree higher than college graduation). As for number of speakers, 46 belong to group 1, 101 to group 2, 104 to group 3, and one speaker is registered once in group 2 and once in group 3.

The last diastratic aspect is speakers' occupation, which is an open category and cannot be treated like the previous ones. Looking at the metadata, the importance of occupations linked to the education field is clear. This happens for different reasons: because professors and students that worked in the corpus compilation are featured in the recordings; because they looked for informants in their social environment (that, of course, is linked to the education system); because age group A is to a great extent formed by students. Nevertheless, in the group linked to education we find students and professors from different faculties, different level teachers, school directors and school clerks. But a significant part of the informants have occupations outside the education system: the corpus features many shop attendants and sales representatives, artists, public clerks, liberal professionals from very different fields (attorneys, doctors, psychologists, dentists, engineers, physiotherapists, etc.), housekeepers, technicians, brokers, craftsman, masons, managers, farmers, and many other occupations.

3.4 Diatopy

As mentioned, the diatopic variation of C-ORAL-BRASIL is essentially that of the Mineiro variety of Brazilian Portuguese (State of Minas Gerais). A corpus of this size must concentrate in representing other variations inside one diatopy. The same happens with the C-ORAL-ROM Project corpora, which represent the regions of Madrid, Marseille, Florence and Lisbon (Cresti & Moneglia 2005). In all the corpora, speakers of other regions and countries are present, since a big metropolitan area comprises a percentage of people from other locations, but what is mandatory for each corpus is that more than 50% directly represent the chosen diatopy. For the C-ORAL-BRASIL this diatopy is the metropolitan area of Belo Horizonte, the state of Minas Gerais capital city. Table 3 shows the informants' origin distribution.

Tab. 3: Speakers' origin.

Origin	Speakers
Belo Horizonte	138
Other cities in Minas Gerais state	89
Other Brazilian states	19
Other countries	2
Unknown	114
TOTAL	362

Excluding the speakers without origin documentation which, as has been already mentioned, account for a non-representative corpus percentage, 55.6% of the speakers are from Belo Horizonte and 35.9% from other cities in Minas Gerais state (many of them from cities within the capital city metropolitan area, like Contagem, Betim, Sete Lagoas, etc.). Therefore, 91.5% of the documented speakers represent the Mineiro variety.

3.5 Transcriptions

An important implementation of C-ORAL-BRASIL was the choice of a specific set of transcription criteria for the segmental part. We wanted to capture a great quantity of phenomena that may be subject to grammaticalization and lexicalization, in order to study them through quantitative methodology and statistic criteria, also measuring their co-occurrence and the systemic relationship among them.

The criteria are based on the following parameters: i. the necessity to represent phenomena subject to grammaticalization and lexicalization (eg.

subject and negation cliticization, loss of verbal morphology, demonstrative reduction, articulated preposition contraction, loss of the verb *ser* in cleft constructions, verb diathesis changes, aphaeresis, and many others); ii. the necessity to keep easy readability of transcriptions, excluding phenomena whose nature was exclusively phonetic, without evident grammatical effect; iii. the necessity to guarantee coherent transcription outputs from transcribers, hence focusing on clearly perceptible phenomena. An example of this last aspect is that of the cliticization of subject pronouns: while the distinction between tonic and clitic forms of the second and third person is relatively easy to perceive (*você(s)* versus *cê(s)* and *ele(s)* or *ela(s)* versus *e ´*, *es*, *ea*, *eas*) the situation is different for the first person singular and plural; in this case we decided not to represent orthographically the opposition between tonic and clitic forms.

All the chosen phenomena are already known by linguists, but they were never documented through corpora. Only spontaneous speech corpus-based studies can truly document: a) how much these and other phenomena are actually recurrent in spontaneous speech; b) to what extent they coexist and determine a deep change in the system; c) which are the most advanced phenomena that may trigger the others; d) what is their distribution based on sociolinguistic variations. If these phenomena were not annotated in the transcription, it would not be possible to study them statistically. In fact, all the forms which differ from the orthographic tradition were implemented in the parser (Bick 2012); this allows a large quantity of studies about on-going linguistic changes that would be impossible to be carried through manual techniques.

4 Conclusion

In this paper I have tried to inspire contact linguists to join forces with corpus linguists in an attempt to broaden the research possibilities related to contact varieties, as well as make them more precise, replicable and potentially incremental through the use of corpora methodologies. In lieu of illustrating the advantages potentially carried by approach based on well architected corpora, the C-ORAL-BRASIL was introduced and its major features were explained. The tradition of exploring actual usage data for contact linguistic analysis can only profit from embracing the rigorous and expanding methodological paradigm brought about by corpus linguistics.

References

Berruto, G. 1987. *Sociolinguistica dell'Italiano Contemporaneo.* Roma: La Nuova Italia Scientifica.

Biber, D. 1988. *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D., S. Conrad & R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, D. & Conrad, S. 2001. Register variation: A corpus approach. In Schiffrin, D. Tannen & H. Hamilton (eds.) *The Handbook of Discourse Analysis*, 175-196. Oxford: Blackwell.

Bick, E. 2012. A anotação gramatical do C-ORAL-BRASIL. In Raso, T. & H. Mello (eds.) C-ORAL-BRASIL I*: Corpus de Referência do Português Brasileiro Informal*, 223-254. Belo Horizonte: Editora UFMG.

Blanche-Benveniste, C., M. Bilger, Ch. Rouget, K. van den Eyde & P. Mertens. 1990. *Le Français Parlé: Études Grammaticales.* Paris: Éditions du CNRS.

Cresti, E. 2000a. *Corpus di Italiano Parlato.* Firenze: Accademia della Crusca.

Cresti, E. 2000b. Critère illocutoire et articulation informative. In Bilger, M. (ed.) *Corpus, méthodologie et applications Linguistiques*, 350-367. Paris: Champion.

Cresti, E. & M. Moneglia. 2005 (eds.) C-ORAL-ROM*: Integrated Reference Corpora For Spoken Romance Languages.* Amsterdam: John Benjamins.

Cresti, E. & M. Moneglia. 2012. In Raso, T. & H. Mello (eds.) C-ORAL-BRASIL *I: Corpus de Referência do Português Brasileiro Falado Informal*, 5-25. Belo Horizonte: Editora UFMG.

De Mauro, T., F. Mancini, M. Vedovelli & M. Voghera. 1993. *Lessico di Frequenza dell'Italiano Parlato.* Milano: ETAS.

Gadet, F. 1996a. Niveaux de langue et variation intrinsèque. *Palympsestes* 10: 17-40.

Gadet, F. 1996b. Variabilité, variation, variété. *Journal of French Language Studies* 1: 75-98.

Gadet, F. 1997 (ed.) *Langue Française* 115. Special Issue on La Variation en Syntaxe.

Gadet, F. 2000. Vers une sociolinguistique des locuteurs. *Sociolinguistica* 14: 99-103.

Gadet, F. 2003. *La Variation Sociale en Français.* Paris: Ophrys.

Hagemeijer, T., I. Hendrickx, H. Amaro & A. Tiny. 2012. A Corpus of Santome. Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012).

Hickey, R. (ed.) 2013. *The Handbook of Language Contact.* Malden, Oxford: Blackwell Publishers.

Holm, J. 2003. *Languages in Contact.The Partial Restructuring of Vernaculars.* Cambridge: Cambridge University Press.

Hornberger, N. 2002. Multilingual Language Policies and the Continua of Biliteracy: an Ecological Approach. *Language Policy* 1: 27-51.

MacWhinney, B. J. 2000. *The* CHILDES *Project: Tools for Analyzing Talk.* 3rd edition. Mahwah: Lawrence Erlbaum Associates.

Mason, M.; Allen, J. 2003. Computing in Creole Languages. *MultiLingual Computing* & Technology 53/14(1). < http://www.multilingual.com/article Detail.php?id=625>

Matras, Y. 2009. *Language Contact.* Cambridge: Cambridge University Press. Mello, H. 1997. *The genesis and development of Brazilian vernacular Portuguese.* Ann Arbor: University Microfilms.

Mello, H. & T. Raso. 2009. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. *Revista Portuguesa de Humanidades* 13: 153-178.

Mello, H., T. Raso, M. Mittmann, H. Vale & P. Côrtes. 2012. Transcrição e segmentação prosodica do Corpus C-ORAL-BRASIL: critérios de implementação e validação. In Raso, T. & H. Mello (eds.) C-ORAL-BRASIL I*: Corpus de Referência do Português Brasileiro Falado Informal*, 125-176. Belo Horizonte: Editora UFMG.

Meyer, C. 2008. Pre-electronic corpora. In Lüdeling, Anke & Merja Kytö (eds.) *Corpus linguistics: an international handbook*, 1-13. Berlin: Walter de Gruyter.

Miller, J. & R. Weinert. 1998. *Spontaneous Spoken Language.* Oxford: Clarendon Press.

Mittmann, M.M. & T. Raso. 2012. The C-ORAL-BRASIL informationally tagged minicorpus. In Mello, H., A. Panunzi, A. & T. Raso (eds.) *Pragmatics and Prosody. Illocution, modality, attitude, Information Patterning and Speech Annotation.* Firenze: FUP.

Moneglia, M. 2005. The C-ORAL-ROM resource. In Cresti, E. & M. Moneglia (eds.) C-ORAL-ROM*: Integrated Reference Corpora for Spoken Romance Languages*, 1-70. Amsterdam: John Benjamins.

Moneglia, M. 2011. Spoken Corpora and Pragmatics. *Revista Brasileira de Linguística Aplicada* 11(2): 479-519.

Moneglia, M. & E. Cresti. 1997. L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In Bortolini, U. & E. Pizzuto (eds.) *Il Progetto* CHILDES *Italia*, 57-90. Pisa: Del Cerro.

Moneglia, M., T. Raso, M. Mittmann & H. Mello. 2010. Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL / C-ORAL-ROM. In Speech Prosody 2010, W1.09. Satellite workshop on Prosodic Prominence: Perceptual, Automatic Identification. Chicago.

Nencioni, G. 1983. *Di scritto e di parlato: discorsi linguistici.* Bologna: Zanichelli.

Raso, T. 2012a. O Corpus C-ORAL-BRASIL. In Raso, T. & H. Mello (eds.) C-ORAL-BRASIL I*: Corpus de Referência do Português Brasileiro Falado Informal*, 55-90. Belo Horizonte: Editora UFMG.

Raso, T. 2012b. O C-ORAL-BRASIL e a Teoria da Língua em Ato. In Raso, T. & H. Mello (eds.) C-ORAL-BRASIL I*: Corpus de Referência do Português Brasileiro Falado Informal*, 91-123. Belo Horizonte: Editora UFMG.

Raso, T., & H. Mello (eds.) 2012 C-ORAL-BRASIL *I: Corpus de Referência do Português Brasileiro Falado Informal.* Belo Horizonte: Editora UFMG.

Raso, T. & M. Mittmann. 2009. Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem* 17(2): 73-91.

Raso, T. & M. Mittmann. 2012. As principais medidas da fala. In Raso, T. & H. Mello (eds.) C-ORAL-BRASIL *I: Corpus de Referência do Português Brasileiro Falado Informal*, 173-221. Belo Horizonte: Editora UFMG.

Robinson, S. 2008. Why pidgin and creole linguistics need the statistician. *Journal of Pidgin and Creole Languages* 23: 141-146.

Scarano, A. (ed.) 2003. *Macro-syntaxe et Pragmatique. L'analyse Linguistique de l' Oral.* Roma: Bulzoni.

Siegel, J., B. Szmrecsanyi & B. Kortmann. In press. Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages.*

Thomason, S. G. 2001. *Language Contact: An Introduction.* Edinburgh: Edinburgh University Press.

Voghera, M. 1992. *Sintassi e Intonazione nell'Italiano Parlato.* Bologna: Il Mulino.

Winford, D. 2003. *An Introduction to Contact Linguistics.* Malden, Oxford: Blackwell Publishers.

Corpora

Brown Corpus (http://khnt.aksis.uib.no/icame/manuals/brown/)

C-ORAL-BRASIL (http://www.c-oral-brasil.org)

Corpus of Northern Haitian Creole (http://www.indiana.edu/˜creole/)

Corpus of Written British (www.ling.lancs.ac.uk/staff/mark/cwbc/cwbcman.htm)

Received: 06/30/2013
Accepted: 12/01/2013